

Proceedings

Open Access

Open Biomedical Ontology-based Medline exploration

Weijian Xuan^{1,4}, Manhong Dai¹, Barbara Mirel^{2,4}, Jean Song^{3,4},
Brian Athey^{1,4}, Stanley J Watson¹ and Fan Meng*^{1,4}

Address: ¹Psychiatry Department and Molecular and Behavioral Neuroscience Institute, University of Michigan, USA, ²School of Education, University of Michigan, USA, ³Health Science Libraries, University of Michigan, USA and ⁴National Center for Integrative Biomedical Informatics, University of Michigan, USA

Email: Weijian Xuan - wxuan@umich.edu; Manhong Dai - daimh@umich.edu; Barbara Mirel - bmirel@umich.edu;
Jean Song - jeansong@umich.edu; Brian Athey - bleu@umich.edu; Stanley J Watson - watsons@umich.edu; Fan Meng* - mengf@umich.edu

* Corresponding author

from Bio-Ontologies 2008: Knowledge in Biology
Toronto, ON, Canada. 20 July 2008

Published: 6 May 2009

BMC Bioinformatics 2009, 10(Suppl 5):S6 doi:10.1186/1471-2105-10-S5-S6

This article is available from: <http://www.biomedcentral.com/1471-2105/10/S5/S6>

© 2009 Xuan et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Effective Medline database exploration is critical for the understanding of high throughput experimental results and the development of novel hypotheses about the mechanisms underlying the targeted biological processes. While existing solutions enhance Medline exploration through different approaches such as document clustering, network presentations of underlying conceptual relationships and the mapping of search results to MeSH and Gene Ontology trees, we believe the use of multiple ontologies from the Open Biomedical Ontology can greatly help researchers to explore literature from different perspectives as well as to quickly locate the most relevant Medline records for further investigation.

Results: We developed an ontology-based interactive Medline exploration solution called PubOnto to enable the interactive exploration and filtering of search results through the use of multiple ontologies from the OBO foundry. The PubOnto program is a rich internet application based on the FLEX platform. It contains a number of interactive tools, visualization capabilities, an open service architecture, and a customizable user interface. It is freely accessible at: <http://brainarray.mbnl.med.umich.edu/brainarray/prototype/pubonto>.

Background

The popularity of data driven biomedical research leads to large volumes of data such as gene expression profiles, MRI images and SNPs related to various pathophysiological processes. As a result, understanding the biological implications of high throughput data has become a major challenge [1]. It requires time-consuming literature and

database mining and is the main goal of "literature-based discovery," "conceptual biology," or more broadly, "electronic biology," which aims at deriving biologically important hypotheses from existing literature and data using *in silico* approaches [2-9]. The effectiveness of such knowledge mining also relies heavily on researchers' back-

ground knowledge about the related data such as novel genes or SNPs, and this knowledge, at present, is sparse.

The Medline database is without doubt the foremost biomedical knowledge database for understanding high throughput data. However, one major shortcoming in prevailing Medline search engines such as PubMed and Google Scholar is that they are designed largely for the efficient retrieval of a small number of records rather than for an in-depth exploration of a large body of literature for discovery and proof purposes. They rely heavily on a step-wise narrowing of search scope [10,11] but such an approach does not work well for the exploration of uncharted territories. In exploration, researchers must be able to apply their background knowledge to define sensible filtering criteria and to infer potentially relevant topics from query results for additional exploration. For example, in microarray gene expression analysis, researchers frequently have to deal with lists of genes that are not known to be associated with the targeted biological processes. Researchers have to utilize other intermediate concepts to establish indirect links between gene lists and specific biological processes. However, identifying such intermediate concepts is very difficult in existing solutions [12], and it is not easy even in systems devoted to this purpose such as ArrowSmith [13,14]. Frequently researchers have to go through large numbers of retrieved records one-by-one and examine external databases to find interesting new relationships. As bioinformatics specialists note, "Continued work is needed to enhance these systems to handle the vast amounts of different types of data that scientists currently must explore manually" [15].

Another shortcoming for prevailing search solutions is that they do not present results in contexts that a user may be interested in. For example, Google Scholar/PubMed presents search results as a linear list of papers, arranged according to citation rate or publication date. Users do not know the context of each paper nor the explicit relationship among the papers. Besides the displayed ranking provided by the search engine, there are few additional cues and sorting/filtering methods that can facilitate the exploration of search results.

We believe the projection of search results to existing knowledge structures is very important for hypothesis development. Cognitive research on scientists' ways of knowing and reasoning shows that to formulate hypotheses from displayed data scientists require cues – data arrangements and conceptual indicators that direct them toward drawing accurate and relevant inferences [16]. These cues determine whether researchers can simply read off lists or can more deeply place retrieved items in a biological context and story [17]. Even if a researcher wants to examine and read off facts in his or her own field from

data displays, there are many details related to the search topic that will still require additional efforts to retrieve, and many of these are not easily identified by prevailing approaches. For example, if Medline search results show that several genes in a brain region are related to a disease in a statistically significant manner, it would be worthwhile to explore the relationship of other genes expressed in this brain region with the disease. Exploring such multiple knowledge structures is often needed to facilitate the formation of new insights. Thus ideally, projecting search results to multiple dynamically-linked knowledge structures could provide the context-assisted data and literature exploration needed for new insights.

Some newer Medline search/mining solutions such as GoPubMed [18], Vivisimo [19], Textpresso [20], and OBIIE [21] attempt to organize search results in the context of either a predefined ontology such as Gene Ontology [22] or dynamically generated ontology structures based on clustering results. Such solutions present a tree-like organization of search results to help users easily navigate to topics of interest. Moreover, the neighbourhood of a given tree branch automatically suggests related topics for additional exploration. Here predefined ontologies have an advantage over clustering results for exploring unfamiliar territories due to their systematic listing and thus cuing of related concepts and their relationships.

These approaches rely on only one or two ontologies for effective exploration, still leaving out or failing to adequately cue many potentially relevant articles previously unknown to the researcher. There are huge numbers of biomedical concepts (e.g., over 1 million in the Unified Medical Language System), and the relationships among them that can be associated with articles are complex. In keeping with scientists' higher order reasoning for exploratory analysis, researchers need to have better tools for formulating searches and examining search results from these different perspectives, including conceptual cues to guide them toward finding biologically meaningful hits. We believe an ontology-based Medline exploration solution must allow the use of different orthogonal ontologies, i.e. ontologies that address different aspects of biomedical research [23]. In addition, for more efficient Medline exploration and for exploration grounded in researchers' domain expertise, it is critical to enable interactive filtering of search results using terms from different ontologies.

The main goal of this work is to develop a flexible ontology-based Medline exploration solution to facilitate the understanding of high throughput data. The solution reveals to researchers potentially interesting conceptual relationships in studies done by other scientists in the target area of investigation. Additionally, the solution lets

researchers selectively view relevant articles by conceptual category within and across multiple levels of different conceptual hierarchies (ontologies); and it lets them interact with visualized ontologies to see search results from diverse perspectives without having to write and re-write query statements. The solution uses multiple ontologies from the OBO foundry. It also has an open architecture that allows flexible selection of Medline retrieval algorithms through different web services.

Methods

Selection of ontologies

The Open Biomedical Ontologies (OBO) foundry is a comprehensive collaborative effort to create controlled vocabularies for shared use across different biological and medical domains [24-26]. It already includes around 50 ontologies from various biomedical domains. We selected Gene Ontology, Foundational Model of Anatomy [27], Mammalian Phenotype Ontology [28], and Environment Ontology [29] for inclusion in our prototype since they provide key perspectives for topics of great interest for many biomedical research and they are almost orthogonal to each other conceptually.

Mapping Medline to ontology

We developed a highly efficient, general purpose ontology to free-text mapping solution in collaboration with researchers in the National Center for Biomedical Ontology. In brief, our solution relies on the pre-generation of lexical variations, word order permutations for ontology terms, their synonyms together with a highly efficient implementation of a suffix-tree based string match algorithm. Our solution is able to map all concepts in UMLS to the full Medline database in 15 hours on a mainstream Opteron server. It achieves over 95% recall rate for UMLS terms when compared to the results from the MMTx program [30], which does not support the use of non-UMLS ontologies. We will describe the technical details and systematic evaluation of our ontology mapping solution in a separate manuscript. Our engine is currently available for web-based OBO annotation for biomedical text at <http://www.bioontology.org/tools/oba.html>.

PubOnto architecture

In order to provide a web-based Medline exploration tool with rich interactivity, we developed PubOnto on Adobe's latest Flex 3.0 platform. It allows us to build a highly interactive user interface that is compatible in virtually all major browsers. We developed an innovative technique that dynamically updates the XML-based ontology tree structure by building a web service for each ontology for feeding expanded nodes with ontological information and literature searching results. As a result, only a minimum amount of data is transferred asynchronously, and PubOnto can thus handle very large ontologies. Figure 1

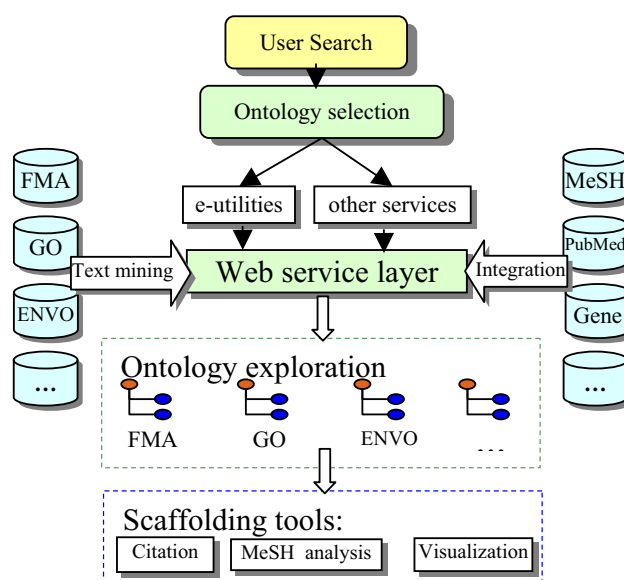


Figure 1
The component-based and open architecture of PubOnto. Technologically, PubOnto is developed based on Adobe's latest Flex 3.0 platform. It allows us to build a highly interactive user interface that is compatible in virtually all major browsers.

shows the architecture of PubOnto. Since the web service layer separates the user interface from ontologies, search services and other databases, the back end changes do not affect the client side user interface.

Results

PubOnto is a FLEX application that provides users with a high level of interactivity for efficient Medline search result exploration. We illustrate a number of key features in this section, tying the features to a use case of exploratory querying and analysis.

Efficient ontology traversal and aggregation

In terms of a use case, a researcher who wants to explore the literature conceptually will enter a query term, e.g. "brain" and then examine retrieved results. He or she will want to see, generally, how retrieved articles are distributed across diverse biological concepts and will want to identify, specifically, which articles deal with the researcher's areas of interest. After submitting the query, the researcher will get a display of results grouped by ontological categories that show how many articles are associated with each superordinate node. He or she will want to know quickly how many articles are retrieved for all children under a branch of any given ontology category – a critical requirement for being able to decide on interesting aspects that need to be explored further. To see these subordinate nodes and to explore the articles associ-

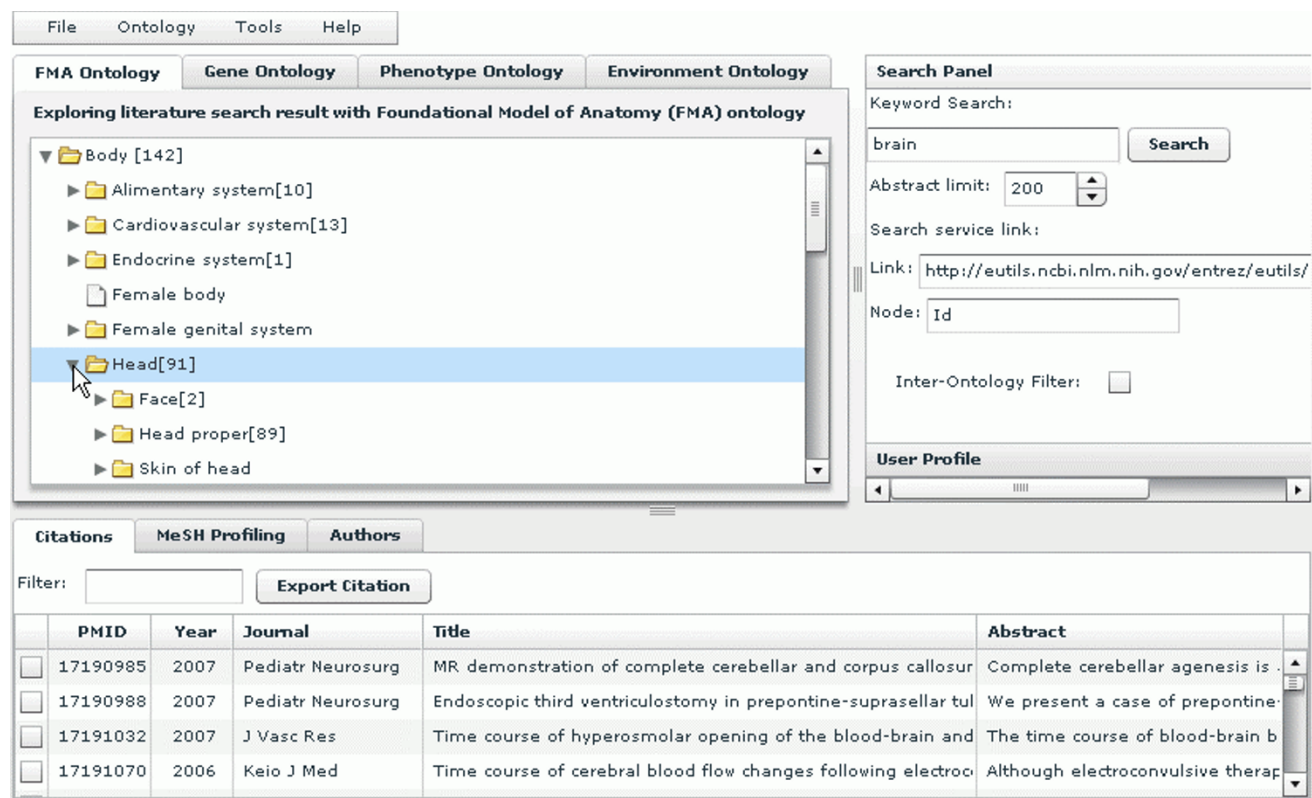


Figure 2
The display of results grouped by ontological categories. It also shows how many articles are retrieved for all children under a branch of any given ontology category.

ated with them, the researcher will expand the tree nodes by clicking on them, as shown in Figure 2; correspondingly, the researcher will collapse tree branches as needed. Expanding and collapsing will be quick and easy.

Technologically, detailing and rolling up such mapped data in a large ontology on-the-fly is not an easy task. Traditional tree traversal algorithms are very CPU intensive and usually require large in-memory tree structures on the server. To provide real time interactivity, we pre-traversed each ontology and generated a parent-child table that matches all nodes in the sub-tree to their parent nodes. When a user issues a query, PubOnto server-side programs save the literature retrieval results to a session-based table. When a user expands a node, our service performs an efficient table join to obtain the aggregated information. This simple and effective algorithm not only provides fast response to user queries but also eliminates the memory pressure on the servers that usually are imposed by ontology traversal and aggregation algorithms that run on-the-fly. This makes PubOnto highly scalable in supporting concurrent users running analyses on multiple ontologies.

Another benefit of this approach is that subsequent exploration of search results on the ontology will be directly returned from a simple database query without any further search, traversal and aggregation. Therefore, from the perspective of a user's experience, PubOnto is highly responsive in interactive explorations.

Ontology-based search result exploration

Continuing with the use case, the biomedical researcher who engages in expanding and collapsing tree nodes in a given ontology will have specific intentions. First, the researcher will want to find concepts of interest and associated articles within the ontology without intervening distractions from the display modes of any other ontologies. Second, the researcher will want to scan and selectively review articles that are associated with both a conceptual category in one ontology and a conceptual category in another ontology. PubOnto features accommodate these different user needs by giving users two search mode options: an independent ontology search mode and an inter-ontology filter mode.

In the independent ontology search mode, when a user clicks the name of an individual ontology concept, the concept becomes highlighted, and citations associated with the concept and its sub-concepts in the current ontology are displayed as a list in the bottom panel (see Figure 2). PubOnto provides an interactive means for expanding parent nodes to their children – namely, users clicking on the arrow next to a concept. It also provides an interactive means for generating a list of associated citations – users clicking on the concept name itself. Numbers in brackets next to a concept name indicate a count of articles for each parent and children node.

In PubOnto's inter-ontology filter mode, when a user clicks the checkbox next to "Inter-ontology filter" (in the lower portion of the Search panel), the search mode for showing intersecting articles between different ontologies is activated (see Figure 3). The inter-ontology filter mode works as follows. In the ontology tab that is open, e.g. FMA, a user specifies a node/categorical concept of interest by clicking on its name. The user then right clicks on

the name of this concept, and a context menu appears. On this menu, the user selects "Map Result to Other Ontologies," as shown in Figure 3. This selection activates PubOnto's automatic mapping of one ontological term to the terms in another ontology. A user sees the results of this mapping by moving to the other ontology. In the other ontology, the count of articles in brackets after the concept name reflects articles belonging to that ontological concept and the the concept mapped from the original ontology. As in the independent search mode, a user can see associated inter-ontology citations by clicking on the name of a mapped concept.

The inter-ontology filter allows a user to easily change concept mappings. If a user returns to the original ontology and chooses a new concept for exploration, he or she can map this new concept to another ontology by selecting "Map Result to Other Ontologies" from the right-click context menu. PubOnto now cancels the previous filtering and maps this newly selected concept to the other

The screenshot displays the PubOnto web application interface. At the top, there are tabs for 'File', 'Ontology', 'Tools', and 'Help'. Below these, the 'FMA Ontology' tab is active, showing a hierarchical tree of anatomical concepts. The 'Gastrointestinal system[50]' is selected, and a right-click context menu is visible with the option 'Map Result to Other Ontologies' highlighted. To the right, the 'Search Panel' includes a 'Keyword Search' field with 'phylogenetic' entered, a 'Search' button, and an 'Inter-Ontology Filter' checkbox that is checked. Below the search panel is a 'User Profile' section. At the bottom, there are tabs for 'Citations', 'MeSH Profiling', and 'Authors'. The 'Citations' tab is active, showing a table of search results.

	PMID	Year	Journal	Title	Abstract
<input type="checkbox"/>	17107469	2006	Mol Ecol	Comprehensive genetic analyses reveal evolutionary distincti	Zapus hudsonius preblei, listed a
<input type="checkbox"/>	17107472	2006	Mol Ecol	Phylogeography of the Western Lyresnake (Trimorphodon bis	The Western Lyresnake (Trimorp
<input type="checkbox"/>	17107556	2006	Environ Microbiol	Study of bacterial communities in Antarctic coastal waters by .	An ecological study on distributio
<input type="checkbox"/>	17107559	2006	Environ Microbiol	A new alpha-proteobacterial clade of Bdellovibrio-like predatc	Bdellovibrio-and-like organisms (

Figure 3

PubOnto's inter-ontology filter mode, which shows intersecting articles between different ontologies. In the ontology tab that is open, a user specifies a node/categorical concept of interest by clicking on its name, and then right clicks on the name of this concept, and a context menu appears. Select the "Map Result to Other Ontologies" activates PubOnto's automatic mapping of one ontological term to the terms in another ontology.

ontologies, thereby filtering to a new set of intersecting articles across ontologies.

This novel functionality provides users both the ontological mapping overview across multiple concept spaces and the detailed results as users navigate through the ontology mapping results. It addresses an important issue that is often encountered by ontology- or clustering-based solutions, which is a lack of interoperability between orthogonal concept spaces.

Ontology selection

Making ontology selections is also part of a typical use case. A researcher, at times, will value exploring results for all the ontologies and, at other times, will want only to see one or two of them. From a use case perspective, rather than "the more function the better," it is "the driving principle is the more suitable the better". PubOnto currently supports four aforementioned OBO ontologies, and with the algorithms and infrastructure implemented, it can easily incorporate a series of many other OBO ontologies. However, in the biomedical research community, researchers usually are specialized in sub-fields such that only a few ontologies are pertinent to them when they perform literature research. Therefore, we present a flexi-

ble way for users to choose which of the supported ontologies they want to use, as shown in Figure 4.

Most existing online literature research tools have rigid user interfaces that do not address heterogeneous user needs. Our approach is different. Ontologies supported by PubOnto are shown in the list. Once a user selects one or more ontologies of interest, PubOnto dynamically creates a new user interface that presents only those ontologies. Each of the ontology tabs has consistent navigational and interactive functions. Search and interaction in PubOnto are conducted only on the selected ontologies. Therefore, this ontology selection feature also reduces the amount of computation required and enhances user interface responsiveness.

Medline citation presentation and analysis

As the use case being presented here suggests, a biomedical researcher will conduct his or her search and analysis tasks by moving back and forth between conducting an analysis of the conceptual distribution of retrieved results and making judgments about the relevance of specific citations associated with conceptual groupings. Seeing details about specific citations will enable the researcher to glean additional information for relevance judgments, for example from the title, author, journal name, date of

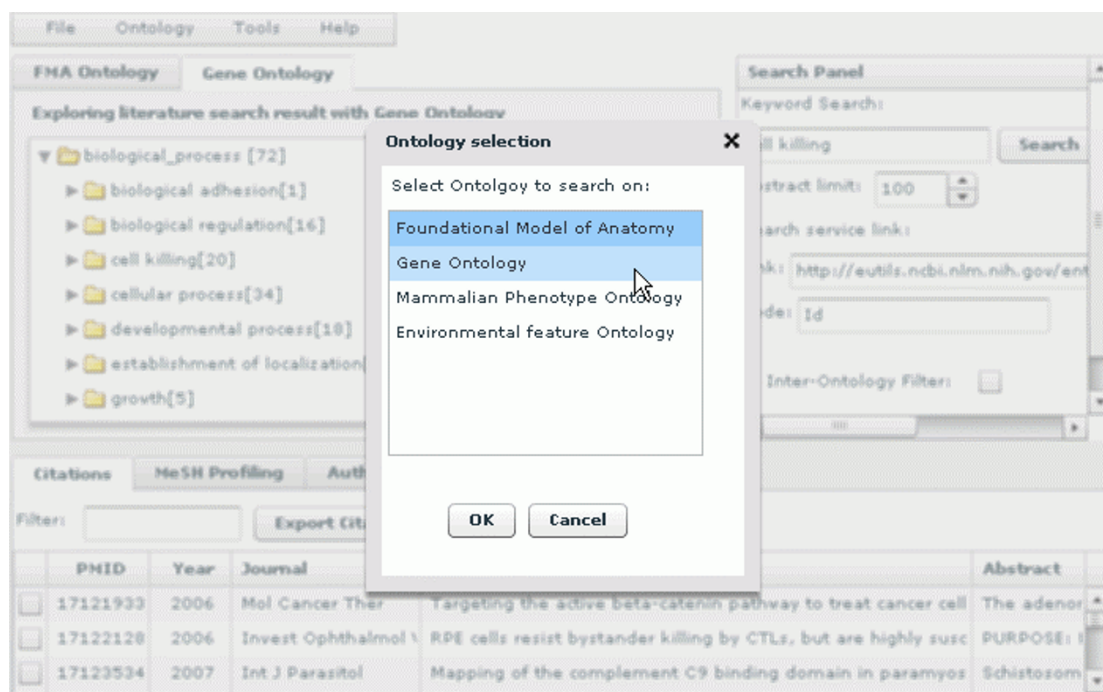
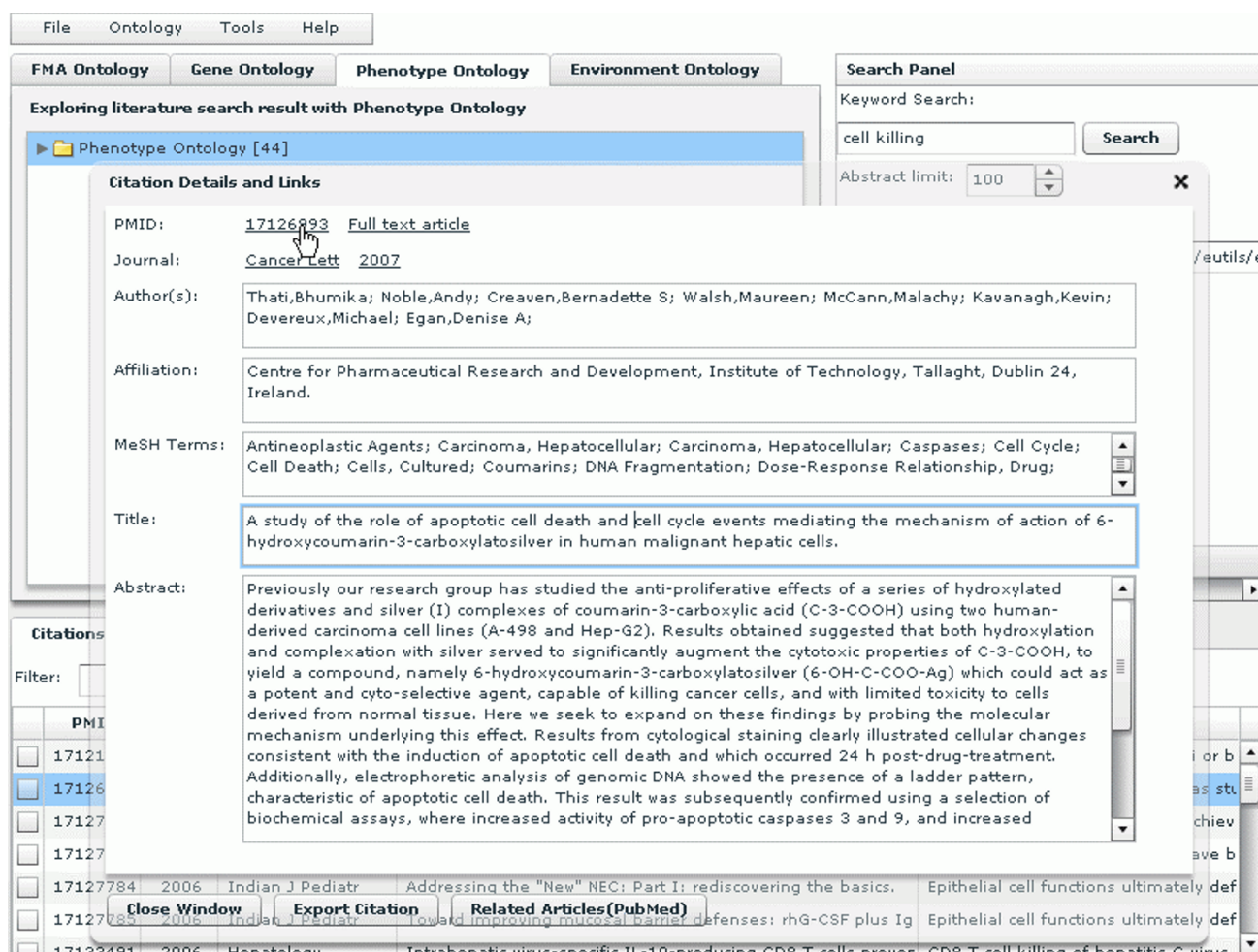


Figure 4

PubOnto supports a flexible way for users to choose which of the supported ontologies they want to use. When a user selects certain ontologies of interest, PubOnto creates a new user interface that presents only the ontologies the user selected.

**Figure 5**

In addition to displaying citations in table format, PubOnto can show citations in detail. To view the detailed record, a user can click the citation to bring up a "Citation Details and Links". It shows Medline record information of this citation, along with links to external resources.

publication, abstract, and associated MeSH terms. Once the researcher finds articles of interest, he or she can export them for future reference.

PubOnto provides users with these export capabilities and with details on citations in the bottom tabbed panel. The panel contains three tabs. The "Citations" tab presents citations associated with an ontology's node (and its children nodes) selected by a user. PubOnto enables a user to view a summary line about each citation or to view in detail the corresponding Medline record. To view the full Medline record, a user can click the citation to bring up a "Citation Details and Links" window, as shown in Figure 5. It shows Medline record information for this citation, along with links to external resources.

PubOnto provides citation filtering capabilities, as well. To filter citations in real time (on the client side) a user can type any keywords in the filter input field, and the citations will be filtered dynamically as users type the keywords. To return to the original citation list, the user just needs to delete the filter string, as shown in Figure 6.

Additionally, because PubOnto is designed to be a Rich Internet Application, it provides users a desktop-like experience. It provides features that enable users to sort citations by any column through a click on the column headers, to re-arrange columns through drag-and-drop header techniques, to resize panels via drag-and-drop panel splitters, and to select multiple citations to process by checking the checkboxes.

Citations					
MeSH Profiling					
Authors					
Filter: nuclear					
Export Citation					
	PMID	Year	Journal	Title	Abstract
<input type="checkbox"/>	17170441	2007	J Gen Virol	Characterization of the IkappaB-like gene family in polydnav	Polydnaviruses (PDVs) are obligate symbionts of hy
<input checked="" type="checkbox"/>	17175527	2007	Mol Biol Evol	The mitochondrial genome of the moss Physcomitrella paten	The phylogenetic positions of bryophytes and charo
<input checked="" type="checkbox"/>	17107469	2006	Mol Ecol	Comprehensive genetic analyses reveal evolutionary distincti	Zapus hudsonius preblei, listed as threatened unde
<input checked="" type="checkbox"/>	17032272	2006	Mol Ecol	Hybridization between mtDNA-defined phylogeographic lineag	Phylogeographic analyses using mitochondrial DNA i
<input type="checkbox"/>	17049280	2007	Mol Phylogen	New World Myotis (Chiroptera, Vesper	Recent studies have shown that species in the genu
<input type="checkbox"/>	17049277	2007	Mol Phylogen	Low nuclear DNA variation supports a recent origin of Hawaiia	Previous phylogenetic work on the Hawaiian bees of
<input checked="" type="checkbox"/>	17101283	2007	Mol Phylogen	Indel evolution of mammalian introns and the utility of non-c	Nuclear DNA intron sequences are increasingly used
<input type="checkbox"/>	17161957	2007	Mol Phylogen	Molecular phylogeny of Clupeiformes (Actinopterygii) inferred	The taxonomy of clupeiforms has been extensively
<input type="checkbox"/>	17185005	2007	Mol Phylogen	A phylogenetic and biogeographic perspective on the evolutio	Phylogenetic relationships of members of the subfa
<input type="checkbox"/>	17116412	2007	Mol Phylogen	Evolution and diversification of the forest and hypogean grou	The beetle genus Trechus (Carabidae) is represent
<input type="checkbox"/>	17188521	2007	Mol Phylogen	Reticulate phylogenetics and phytogeographical structure of t	The Balkan Peninsula is known to be one of the mo
<input type="checkbox"/>	17178747	2007	Nucleic Acids Res	An enhanced MITOMAP with a global mtDNA mutational phylo	The MITOMAP (http://www.mitomap.org) data syste
<input type="checkbox"/>	17144892	2007	Plant J	The Arabidopsis ClpB/Hsp100 family of proteins: chaperones	The Casein lytic proteinase/heat shock protein 100

Figure 6

PubOnto provides citation filtering capabilities. To filter citations on the client side, a user can type any keywords in the filter input field, and the citations will be filtered dynamically as users type the keywords. To return to the original citation list, the user just needs to delete the filter string.

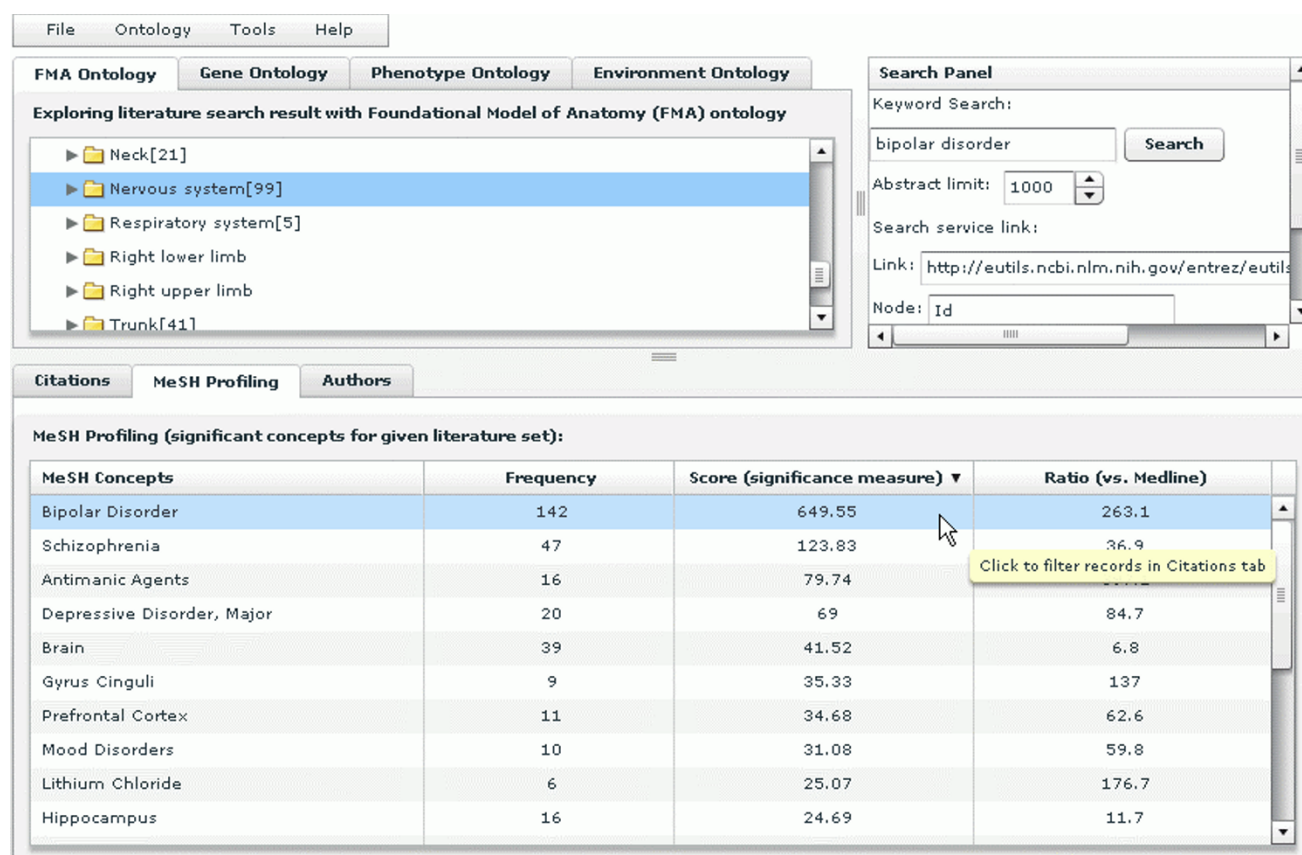
PubOnto also satisfies users' needs to export citations of interest. When users choose "Export Citation" PubOnto will format citation information, automatically launch Endnote installed on the client side, and import the information to the citation library that the user specifies.

In another tab in this lower panel, "MeSH Profiling," the aim is to help users identify themes and contexts of the publications to facilitate users in exploring potentially related publications. The "MeSH Profiling" function identifies the most significant MeSH concepts that differentiate the current citation group from the whole Medline corpus. We adapted the corpus profiling algorithm [31], which ranks terms that differentiate two corpora the most. In PubOnto, we always derive the rank of terms based on the retrieved records vs. all Medline records. For example, if a user queries for "bipolar disorder" and selects "Nervous system" from the FMA ontology, the MeSH concept "Bipolar Disorder" is ranked the first by the algorithm, and other related concepts such as disorders, brain structures and drugs also rank among the top. These concepts should be very useful for researchers not familiar with this area to conduct further exploration for hypothesis development (Figure 7). Because we have pre-calculated the frequency of each MeSH term in Medline, the profiling calculation only takes less than half a second. Simple frequency counts and ratios of the frequency of a term in the current group to the whole Medline corpus are also provided so that users can sort the MeSH concepts by any of these criteria.

In the third tab, "Authors," PubOnto offers a summary of the authors of the current citation group. A user can sort by citation count to find authors published frequently on this topic or follow the links in the table to find other publications of an author. Meanwhile, with the same techniques and underlying reasoning discussed in the previous section, PubOnto allows a user to customize display of the data analysis tabs.

Data visualization

Data visualizations are relevant to our exploratory search and analysis use case, as well. A biomedical researcher will want to identify quickly which ontological categories in one of more ontology contain retrieved articles. He or she also will want to conceptualize how the articles overall are distributed across these ontological categories and what trends characterize publication dates of articles in various conceptual categories. Grasping these "results profiles" will help direct the researcher to the concepts and articles most pertinent to his or her goals. We provide visual presentations to enable analysts to quickly grasp frequencies, distributions, and trends [32]. PubOnto has separate graphic tools to facilitate users in discerning from the visualized structure of retrieved results the spread of concepts relevant to the query terms or the apportioning of retrieved articles in these concepts. To graphically display frequencies, distributions, and trends PubOnto provides four types of common charts, including a pie, line, column, and plot chart (Figure 8a). Each chart displays the data suited to a given type of analysis. For example, a user can pick a pie chart to visualize the proportionate spread

**Figure 7**

MeSH profiling function aims to help users identify themes and contexts of the publications so that they can better identify those of interest. It identifies the most significant MeSH concepts that differentiate the current citation group from the whole Medline corpus.

of journals (pie slices) in a given result set. This view immediately shows users which domain specialities devote more coverage to the topic being searched than others. See Figure 8 for the procedures and outcomes of displaying these PubOnto visualization tools.

All charts are interactive. For example, using the PubOnto mouse over function a user can view more details of a graph. A user also can go back to previous selections to change data sources or chart types (Figure 8d). Note that for some charting, PubOnto automatically performs and displays aggregations after a user specifies the chart type and data source. For other charts PubOnto uses the same data displayed in the tables. Yet visual displays present results more intuitively for interpretation.

Search service and open architecture

Information retrieval algorithms in the biomedical fields often have pros and cons but existing Medline search solutions do not take this into consideration. We believe one search algorithm does not fit all needs in diversified bio-

medical fields. Thus we designed an open architecture that can incorporate other search algorithms through web services in addition to the default search service. In the default search service, when users submit a query the client-side program sends web service requests to service programs on the server. These service programs perform the following tasks: 1) transform the query and other criteria, 2) call search services provided by PubMed, 3) extract PMIDs, 4) map to pre-indexed Medline-to-ontology mapping tables, 5) aggregate results and handles potential search exceptions, and 6) finally return the results in web service format to the client program.

In PubOnto, the service is designed in a component-based architecture. Such a design enables us to open the tasks 2) and 3) above to other search services that can be specified by users (where the users also need to specify the element that contains PMID). Once the result is retrieved the PMIDs are extracted from a third-party web service and the subsequent tasks are performed as in the default mode. In either mode, the search service is controlled by



Page 10 of 14

our service program instead of directly interacting with the client side program. The operations are transparent to users without problems related to security or cross-domain accessibility. The only change, which is also the purpose of this architecture, is the difference in search results that reply on the search service the user specified.

Other scaffolding tools

From a use case perspective, a researcher expects searching, creating filters, and navigating to be easy and efficient. PubOnto offers several tools for these purposes, including an auto-complete function for keyword search, customizing features in the context menu mentioned previously, and user session management.

1) Auto-Complete function. This function facilitates typing in search keywords in the input box, thereby making users' formulations of queries efficient. As a user types, PubOnto make a service call, which matches the string to entities in ontology databases and returns a list of partially or fully matched terms. While somewhat sophisticated in implementation, it is intuitive to users (it behaves the same as other popular search engines do, e.g., Google's auto-complete function.).

2) Customizable context menu. During the ontology-based exploration, users may find literature hits pointing to concepts that are unfamiliar to them. In such situations, a user will want to quickly examine the meaning of the concept or search for the context in which the concept is used. Typically, a user will want to turn to his or her favourite search engine, and PubOnto accommodates this intention. PubOnto implements a context-sensitive menu that can be accessed by a right click on the ontology node. Existing options in the menu include getting term definitions from our in-house integrated ontology database and directing to specific entries in Wikipedia or PubMed. A user can add other search engines or other services, as shown in Figure 9. A user can also decide what menu items the context menu entry should have. The customized context-sensitive choices behave the same way as built in options.

3) User session management. The interactive exploration capability of PubOnto enables users to search and filter results in more combinations than typical tools provide but it also makes it harder to trace back the processes. PubOnto thus builds in a user session management function. It records users' search history and their choice of parameters (Figure 10) so that users can save or trace their search sessions.

In summary, while the most important feature of PubOnto is the ability to use multiple OBO ontologies for Medline exploration, it also offers a number of unique

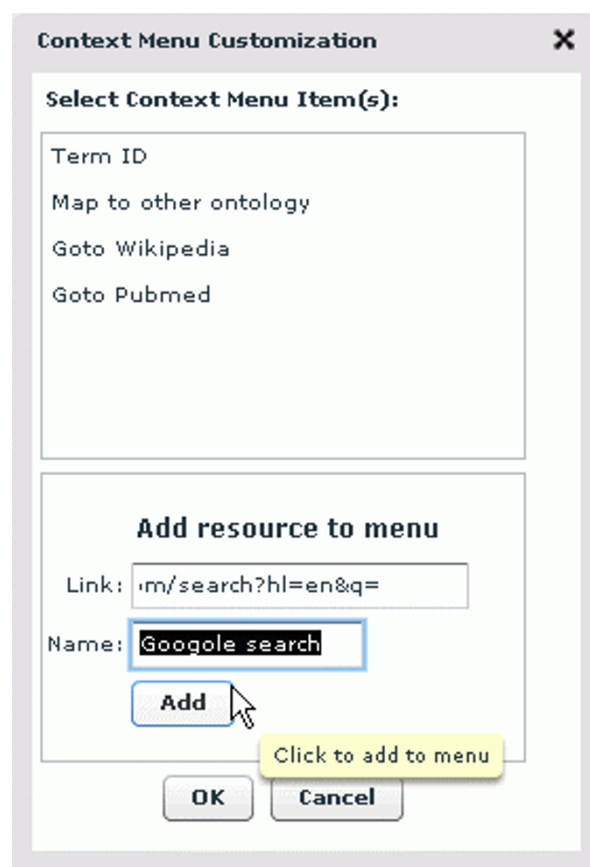
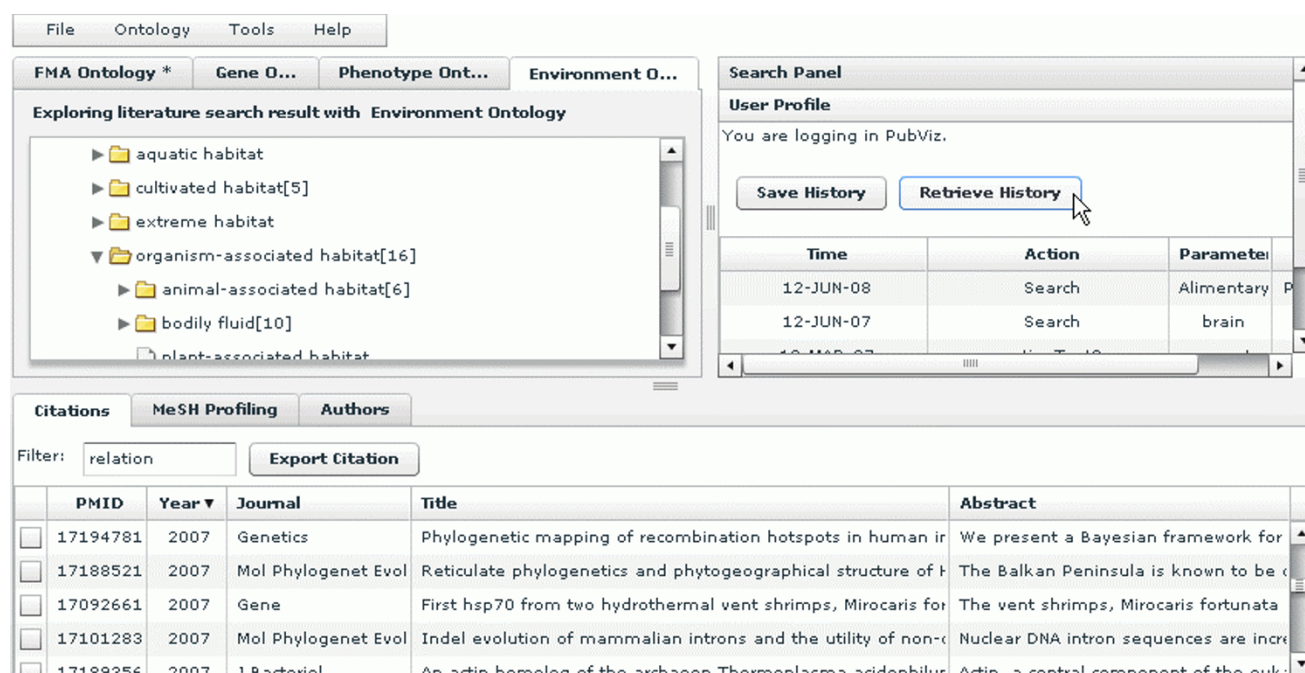


Figure 9
PubOnto builds in a context-sensitive menu that can be accessed by a right click on the ontology node. In addition to the default options on the context menu, users can add other search engines or other services. The customized context-sensitive choices behave the same way as built in options.

features summarized in Table 1. Table 1 compares our solution with GoPubMed, which is arguably the most popular ontology-based literature search program in the biomedical field. The PubOnto prototype currently does not include several functions that GoPubMed has, as they are not directly related to ontology, but similar functions will be added in the future upon users' request.

Discussion

Systematic ontology development efforts such as those related to the Open Biomedical Ontologies are likely to generate an expansive conceptual framework for the integration, analysis and understanding of data generated in different areas of biomedical research. PubOnto aims to capitalize on the impressive progress in ontology development for the exploration and mining of biomedical literature. The ability to utilize multiple orthogonal

**Figure 10**

User session management. PubOnto prototyped a user session management function. It records users' search history and their choice of parameters. After logging in, users can save or trace their search sessions.

ontologies during Medline exploration can significantly increase the efficiency of locating interesting search results in areas that researchers are not familiar with. Mapping Medline results to multiple ontologies also enables researchers to explore search results from different angles for new hypothesis development.

While PubOnto enables the use of multiple ontologies for Medline exploration, there are a number of improvements

Table 1: Comparison between PubOnto and GoPubMed

	PubOnto	GoPubMed
More ontologies besides MeSH, GO	Yes	No
Interaction among ontology	Yes	No
Customizable search service	Yes	No
Client side filter	Yes	No
Customizable ontology search	Yes	No
Customizable interaction function	Yes	No
Rich interactions	Yes	No
Search history maintenance	Yes	Yes
Sorting citation by various criteria	Yes	No
Export to Citation managers	Direct	Indirect
Citation linkouts	Yes	Yes
Where/Where/When analysis	Some	Yes
Keyword highlight	No	Yes
Hot topics	No	Yes
Wikipedia mapping	Yes	Yes

we hope to incorporate in future versions. For example, although the ability to select different ontologies for organizing search results is quite powerful, it is based on the assumption that users know which ontologies they want to use. In reality, due to the quick expansion of OBO, a typical user may not know the most useful ontologies for his/her literature exploration beforehand. It will be ideal for PubOnto to provide some automated ontology selections as the starting points of exploration. We plan to develop methods to rank ontologies for their usefulness to a specific topic based on the distribution of returned Medline records on different concepts under a given ontology. For example, an ontology is not very useful for Medline search result exploration if only a small fraction of returned records can be mapped to this ontology. On the contrary, an ontology will be very effective if many records can be mapped to it and if those records are relatively evenly distributed across many terms in that ontology. Of course, an ontology will still not be useful if most of the search results can be mapped to only a few terms in an ontology. Consequently, it should be possible to develop an ontology scoring system based on the number of records that can be mapped to an ontology and the distribution of Medline records in an ontology for the automatic selection of a default ontology for a given Medline search result. Conceivably, once the first ontology is selected, it is possible to select the second best

ontology based on the "orthogonality" with the first ontology. Of course, such automated ontology ranking procedures are only based on the statistical properties of the Medline records to ontology mapping. Users' biomedical knowledge and their understanding of different ontologies will be essential for effective exploration of Medline literature.

Similarly, the exploration of a given ontology tree currently is also dependent on users' background knowledge since only the number of Medline records hits for a given term can be used as external cues for ontology exploration now. If there are many different ontologies from which a user is to select or if a user is not familiar with the corresponding ontology at all, it is desirable to have additional information to help users engage ontologies for exploration more effectively and efficiently. It will be pretty straightforward to weigh the specificity of each ontology term based on their inverse frequency of showing up in the Medline corpus so that users can focus on more specific terms rather than exploring generic terms.

Another improvement involves web services. PubOnto is designed with a client-server model, and all the communications between the client and server are through standard web service or http service calls. As we have done for other programs, we will open our ontology-based search and analysis web services to the community to help the development of other research projects.

These technological advances will be coupled with user testing. As Cohen and Hersh [15] highlighted, many retrieval and analysis tools in bioinformatics fall short in supporting scientists' actual exploratory analysis. We will use formative usability testing to assure that designs of current and new features and displays adequately match the purposes and practices of scientific researchers. Testing will uncover, as well, necessary improvements for ease and efficiency of use, navigation, comprehension and satisfaction. After fitness-to-purpose and usability improvements are implemented, we will conduct comparative user performance evaluations with systems such as PubMed and GoPubMed with the aim of establishing the benefits of PubOnto's technological advances and user-centered design.

Conclusion

We believe the use of multiple ontologies in OBO for Medline exploration can significantly increase the efficiency of Medline exploration and facilitate the examination of the same search result from different perspectives. We will continue to improve PubOnto to make it an effective tool for novel biomedical hypotheses development, and ultimately incorporate it into PubViz, our more comprehensive biomedical literature exploration engine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

W. Xuan is the main developer of the PubOnto project. M. Dai developed the free text to Ontology mapping engine used in this project. All authors participated in the design, evaluation and improvement of the PubOnto with main efforts from W. Xuan, B. Mirel and F. Meng. All authors contributed to and approved the manuscript.

Acknowledgements

W. Xuan, M. Dai, S. J. Watson and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through NIH grant 1U54DA021519-01A1 to the University of Michigan.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 5, 2009: Proceedings of the Bio-Ontologies Special Interest Group Workshop 2008: Knowledge in Biology. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/10?issue=S5>.

References

1. Boguski MS, McIntosh MW: **Biomedical informatics for proteomics.** *Nature* 2003, **422**(6928):233-237.
2. Swanson DR: **Medical literature as a potential source of new knowledge.** *Bull Med Libr Assoc* 1990, **78**(1):29-37.
3. Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery.** *Nature Reviews Genetics* 2006, **7**:119-129.
4. Srinivasan P: **Text mining: Generating hypotheses from MEDLINE.** *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY* 2004, **55**(5):396-413.
5. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR: **Knowledge discovery by automated identification and ranking of implicit relationships.** *Bioinformatics* 2004, **20**(3):389-398.
6. Kim J-D, Ohta T, Tsujii J: **Corpus annotation for mining biomedical events from literature.** *BMC Bioinformatics* 2008, **9**:10.
7. Hu X, Wu D: **Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases.** *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2007, **4**(2):251-263.
8. Roberts PM: **Mining literature for systems biology.** *Brief Bioinform* 2006, **7**:399-406.
9. Erhardt RA-A, Schneider R, Blaschke C: **Status of text-mining techniques applied to biomedical text.** *Drug Discovery Today* 2006, **11**(7-8):315-325.
10. Hoogendam A, Stalenhoef A, Robbe P, Overbeek AJ: **Analysis of queries sent to PubMed at the point of care: Observation of search behaviour in a medical teaching hospital.** *BMC Med Inform Decis Mak* 2008, **8**:42.
11. Herskovic JR, Tantaka LY, Hersh W, Bernstam EV: **A day in the life of PubMed: Analysis of a typical day's query log.** *J Am Med Inform Assoc* 2007, **14**(2):212-220.
12. Chen H, Sharp B: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5**(1):147.
13. Smalheiser NR, Torvik VI, Lugli G, Zhang W, Zhou W, Hulth M, West R, Weeber M, Swanson DR, Maryann Martone P, et al.: **Arrowsmith.** 2007 [http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html].
14. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med* 1986, **30**(1):7-18.
15. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Brief Bioinform* 2005, **6**(1):57-71.
16. Carruthers P, Sigal M, Stice S: **The Cognitive Basis of Science.** Cambridge: Cambridge University Press; 2002.

17. Kuchinsky A, Graham K, Moh D, Adler A, Barbaria K, Creech M: **Biological storytelling: A software tool for biological information organization based on narrative structure.** *Proceedings of ACM-SIGGROUP AVI'00, 2000: 2000; Trento, Italy 2000.*
18. Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology.** *Nucleic Acids Res* 2005:V783-786.
19. Taylor DP: **An integrated biomedical knowledge extraction and analysis platform: using federated search and document clustering technology.** *Methods Mol Biol* 2007, **356**:293-300.
20. Muller H-M, Kenny EE, Sternberg PV: **Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature.** *PLoS Biology* 2004, **2(11)**:e309.
21. Milward D, Bjärelund M, Hayes W, Maxwell M, Oberg L, Tilford N, Thomas J, Hale R, Knight S, Barnes J: **Ontology-Based Interactive Information Extraction from Scientific Abstracts.** *Proc of BioLINK SIG Text Mining Workshop: 2004; Glasgow 2004.*
22. Consortium TGO: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25-29.
23. Shimoyama M, Petri V, Pasko D, Bromberg S, Wu WH, Chen JL, Nenashina N, Kwitek A, Twigger S, Jacob H: **Using multiple ontologies to integrate complex biological data.** *Comp Funct Genom* 2005, **6(7-8)**:373-378.
24. NCBO: **The Open Biomedical Ontologies.** 2008 [<http://www.obofoundry.org/>].
25. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al.: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, **25(11)**:1251-1255.
26. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA, et al.: **National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge.** *Omics* 2006, **10(2)**:185-198.
27. Rosse C, Mejino JL: **A reference ontology for biomedical informatics: the Foundational Model of Anatomy.** *J Biomed Inform* 2003, **36(6)**:478-500.
28. Smith C, Goldsmith C-A, Eppig J: **The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.** *Genome Biology* 2004, **6(1)**:R7.
29. Smith B, Kusnierczyk W, Schober D, Ceusters VV: **Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain.** *Proceedings of KR-MED: 2006* 2006:57-66.
30. Aronson AR: **Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap Program.** *Proc AMIA Symp* 2001:17-21.
31. Rayson P, Garside R: **Comparing corpora using frequency profiling.** *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics: 2000; Hong Kong 2000*:1-6.
32. Yi JS, Kang Y-a, Stasko JT, Jacko JA: **Understanding and characterizing insights: how do people gain insights using information visualization?** In *Proceedings of the 2008 conference on Beyond time and errors: novel evaluation methods for Information Visualization: 2008; Florence, Italy ACM; 2008*:1-6.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

